DOCUMENT RESUME

ED 312 083 PS 018 369

AUTHOR Lichtenstein, Robert
TITLE Psychometric Characteristics of the Gesell School Readiness Screening Test.
PUB DATE 88
NOTE 29p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Concurrent Validity; Kindergarten; *Kindergarten Children; Primary Education; *School Readiness; *Screening Tests; Student Placement; *Test Reliability; *Test Use; *Test Validity
IDENTIFIERS *Gesell School Readiness Screening Test; Test Retest Reliability

ABSTRACT

The Gesell School Readiness Screening Test (GSRST) is widely used to identify "developmentally immature" children for placement in extra-year, transition programs in spite of a problematic absence of psychometric evidence and research support. In this study of psychometric characteristics of the GSRST, teacher ratings of classroom performance and readiness, selected subtests from the Kaufman Assessment Battery for Children (K-ABC), and parent reports of adaptive behavior were obtained for a sample of 48 kindergarten children. There was a moderately high correlation between the GSRST and K-ABC subtests. There was a high level of agreement between placement recommendations based on GSRST results and teacher ratings of readiness. Test-retest reliability and interrater reliability for the customary clinically-based scoring method were below acceptable levels, and lower than the reliability figures yielded by a quantitatively based scoring method. The GSRST does not appear to be sufficiently reliable for use as a primary determinant of grade placement, but may be a valuable component of a school readiness assessment. The construct of school readiness as measured by the GSRST remains in need of clearer definition. (Author/RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED312083

Psychometric Characteristics of the Gesell School Readiness Screening Test

Robert Lichtenstein

Gesell Institute of Human Development

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Robert Lichtenstein
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The participating educators, parents, and schoolchildren are gratefully acknowledged for making this study possible. Thanks also to Dr. George McCloskey of American Guidance Service for the loan of assessment materials for use in this study.

The author is no longer affiliated with Gesell Institute of Human Development. Correspondence with the author should be addressed to 896 Quinnipiac Avenue #3, New Haven, CT, 06513.

PS 018369

BEST COPY AVAILABLE

Abstract

Despite the problematic absence of psychometric evidence and research support, the Gesell School Readiness Screening Test (GSRST) is widely used to identify "developmentally immature" children for placement in extra year (eg., transition) programs. Teacher ratings of classroom performance and readiness, selected subtests from the Kaufman Assessment Battery for Children (K-ABC), and parent reports of adaptive behavior were obtained for a sample of 48 kindergarten children in order to investigate psychometric characteristics of the GSRST. Test-retest reliability was also assessed. A moderately high correlation (r = .59) was found between the GSRST and K-ABC subtests. A high level of agreement (81 percent) was found between placement recommendations based on GSRST results and teacher ratings of readiness, although both identified children as unready for first grade at a very high rate (58 percent and 69 percent, respectively). Test-retest reliability and inter-rater reliability for the customary clinically-based scoring method were below acceptable levels, and lower than the reliability figures yielded by a quantitatively-based scoring method. The GSRST does not appear to be sufficiently reliable for use as a primary determinant of grade placement, but may be a valuable component of a school readiness assessment.

Psychometric Characteristics of the Gesell School Readiness Screening Test

Assessing for school readiness has become a practice of considerable interest and concern. Many school districts routinely screen all children before or during kindergarten for school readiness, placing children who test as "unready" or "developmentally immature" in extra year programs. Thus, a child who is judged unready for first grade may be placed in a "pre-first grade transition" or "readiness" classroom for an extra year before entering first grade, and a child judged unready for kindergarten may be recommended for a "developmental kindergarten" or "prekindergarten" class.

The practice is widespread: a report by Educational Research Service (1986) found that 17.5 percent of the school districts surveyed offer a transitional room program for students not fully ready for first grade--and, that 87.8 percent of the districts conduct standardized testing in kindergarten! The practice is also controversial: parents of a five-year-old in upstate New York received a favorable settlement of their legal complaint directed at the school district's policy of relying on test results to assign children to developmental kindergarten.

The issue of whether extra year programs offer real and appreciable benefits is a troubling one. While strongly-worded position statements by professional organizations (eg., National Association for the Education of Young Children, 1987; National Association of Early Childhood Specialists in State Departments of Education, 1987) alert us to the fact that no benefits have been empirically demonstrated, one can argue in turn that no published research clearly indicts the practice. The two research references commonly cited in the case against extra year programs are a

study by Shepard and Smith (1985) that fails to differentiate between grade retention and placement in extra year programs and a review paper by Gredler (1984) that summarizes findings from unpublished research studies.

Regardless of how one interprets the research to date, critics draw attention to aspects of extra year programs that appear problematic because of measurement concerns. First, it is questionable whether children who are developmentally immature, or lacking in "school readiness," can be identified with sufficient accuracy to justify making decisions that will result in an extra year of school for the child (Meisels, 1987, in press; Peck, McCaig, & Sapp, 1988; Shepard & Smith, 1986). Second, if one accepts the premise that readiness tests measure essentially the same thing as intelligence tests, placing children in classrooms according to school readiness is tantamount to tracking by ability level (Shepard & Smith, 1986). A related assumption is that , as a consequence of this, low-functioning children and those who test poorly (eg., minority children) will be unnecessarily exposed to restrictive classroom environments, lowered performance expectations, and the negative stigma of special class assignment (National Association of Early Childhood Specialists in State Departments of Education, 1987).

The Gesell School Readiness Screening Test (Ilg & Ames, 1964) and related tests adapted from Arnold Gesell's research (Gesell Preschool Test, Gesell School Readiness Test) are the most widely used measures for assessing school readiness. This is not surprising given that they are the only published measures explicitly designed to identify children who are functioning within the normal range but developmentally unready for a given grade or educational environment.

The Gesell School Readiness Screening Test (GSRST), the Gesell test most commonly used to screen for extra year placement, is interpreted in a manner that is uncharacteristic of individually-administered standardized tests. The test is designed

to elicit a sample of responses and behaviors that are observed by a developmental examiner who has been extensively trained in the use of the test. Rather than objectively score and tally test items, examiners clinically interpret a child's behavior relative to their knowledge of developmental expectations. As explained in the administration manual, "One considers the processes, the organization, the method, the overt behaviors, and verbalization, as well as the end product, to determine the child's overall developmental age." (Gesell Institute, 1985, p. 2).

Developmental age is regarded as "the age at which the child is functioning as a total organism--the social, emotional, intellectual and physical components are interdependent" (Gesell Institute, 1985, p. 1). Developmental age translates into educational recommendations in that children falling below the cutoff developmental age at a given point in the school year are seen as unready to progress to the next grade. Despite the familiar caveat to integrate test findings with other sources of information in evaluating children's developmental behavior (eg., California State Department of Education, 1988; Lichtenstein & Ireton, 1984; National Association for the Education of Young Children, 1987), many school districts rely upon the Gesell test alone to make grade placement recommendations about individual children.

Meisels (1987), in writing about uses and abuses of Gesell testing, states that, "no systematic study of the validity of these tests has ever been conducted." The statement is technically incorrect: reliability and validity data on the 1970's version of the Gesell Preschool Test have now been prepared for publication (Walker, 1988), and Kaufman (1971) conducted psychometric analyses of a quantitatively-scored research version of the Gesell test which was validated against subsequent academic performance (Kaufman & Kaufman, 1972). Meisels and others (Bear & Modlin, 1987; Bradley, 1985; Kaufman, 1985) are warranted, however, in calling attention to the woefully inadequate psychometric underpinnings of the Gesell assessments. There is

no evidence of any reliability or validity investigations having contributed to the development of the Gesell test, and the normative data reported (Ilg & Ames, 1964) are not of a conventional nature. Furthermore, statements concerning appropriate use of the GSRST are conspicuously absent from the published version of the test and from the scoring and administration guide distributed by Gesell Institute (1985) for the training of developmental examiners.

In the absence of supportive psychometric evidence, users of the GSRST typically state that they believe the test is accurate because it corroborates the judgement of the classroom teacher. This, clearly, is a premise that merits empirical investigation. The present study addresses this and other psychometric issues: (1) Does the GSRST measure "developmental maturation" as intended, as opposed to cognitive ability? (2) How accurately does it identify children who are unready to progress to the next grade level? (3) Is it culturally fair, ie., does it yield placement recommendations that are unbiased with respect to different demographic subgroups? (4) Is the clinically-based scoring method sufficiently reliable to justify interpreting the test in this manner and using it to make placement recommendations?

These measurement issues impact upon the larger question of the value of extra year programs. If children are being assigned to extra year programs on the basis of unsatisfactory measures or inappropriate criteria., this will confound efforts to conduct a fair and rigorous assessment of the effects of these programs.

Method

Subjects

A battery of development screening measures was administered to a sample of 48 kindergarten children, 24 males and 24 females, in an ethnically diverse community in an outlying part of the metropolitan New York City area. At the time the full

battery was administered, the sample ranged in age from 64 to 84 months, with a mean of 72.2 months and standard deviation of 5.1 months. The ethnic composition of the sample, which was representative of the overall kindergarten class and school district, was 54.2 percent white, 33.3 percent Hispanic, 10.4 percent black, and 2.1 percent Asian-American. Socioeconomic status (SES) of the children's families were distributed fairly evenly from lower-middle to upper-middle class range. All children in the sample attended the same school, which houses all of the district's kindergartens.

Procedure

Children in the study were administered the GSRST on two occasions during the last two months of the school year in order to assess test-retest reliability. The first administration was conducted during the school day by school district personnel. For the second test administration, parents brought their children to the school for a special testing session in which a battery of developmental screening information was obtained. Parents provided background information about the child and completed a standardized inventory of the child's adaptive behavior, the Self-Help/Social Questionnaire of the Early Screening Profiles (Sparrow & Cicchetti, in press). In addition, children were administered four representative subtests--Gestalt Closure, Number Recall, Triangles, and Riddles--from the Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983), a measure of general cognitive ability standardized on a nationally representative sample. Test batteries were administered by eight trained examiners, of which four were from the school district and four from nearby school districts. Two of the examiners were kindergarten teachers at the school. The retest examiner was never the initial examiner, nor was either the child's classroom teacher. All examiners in this study were certified by Gesell Institute as "qualified examiners," the highest level of certification accorded. For comparability

across measures, GSRST results from the second testing session are reported here when not otherwise specified.

Kindergarten teachers provided the criterion measure of whether children were developmentally ready for first grade. Teachers made judgements about the optimal placement of each child for the following school year, selecting from among the following choices: (1) Child should definitely be ready for first grade, (2) Child may not be ready for first grade, (3) Child will definitely not be ready for first grade, (4) Child will probably require special educational services. Teachers were also asked to rate all of the children in their classrooms on six global dimensions--learning habits, social maturity, language skills, gross motor skills, pre-academic skills, and perceptual/fine motor skills--using a seven-point scale. A total of 182 children in nine classrooms, six regular kindergartens and three developmental kindergartens, were rated by their teachers.

Results

Descriptive statistics for the GSRST and for the composite of K-ABC subtests are shown in Table 1. Gesell developmental age (DA) scores, usually expressed as half-year

---

Insert Table 1 about here

---

intervals and combinations thereof, were translated into a continuous scale with two-month intervals for analyses requiring an interval scale. A score of "5 / 5-1/2" ("five, five-and-a-half") was regarded as a DA of 62 months, a score of "5-1/2 / 5" was regarded as 64 months, and so on. Age-adjusted scores are presented in Table 1: developmental age/chronological age (DA/CA) discrepancy for the GSRST and mean

subtest scaled score (based on a mean of 10 and standard deviation of 3) for the K-ABC composite.

The distribution of Gesell DA scores was highly skewed. Although half of the sample was six years of age or older, the highest score was a DA of 72 months, obtained by a single child. Consequently, Gesell DA/CA discrepancy scores were markedly greater for children six years and older than for five year olds, and the older half of the sample all had negative DA/CA discrepancy scores. Accordingly, there was a strong negative correlation between Gesell DA/CA discrepancy and chronological age (r = -.54). The sample obtained a mean K-ABC composite scaled score of 9.9, which translates into an IQ equivalent of 98.5. Females scored lower than males on the K-ABC, but slightly higher than males on the GSRST, with considerably less variance in performance on both measures. Expected subgroup differences by SES (lower-middle vs. middle/upper-middle), language status (primarily English-speaking vs. bilingual), and ethnic background were found for K-ABC performance, but not for GSRST performance.

Although it would have been preferable to conduct correlational analyses using K-ABC standard scores and Gesell DA/CA discrepancy scores so as to eliminate age effects, this was inadvisable given the ceiling effect just described. The variables Gesell DA and K-ABC composite raw score were used instead. Gesell DA correlated .32 with chronological age, in comparison with .19 for the K-ABC composite. These figures should be regarded in the context of a slightly reduced relationship between age and ability for the sample, as evidenced by the negative correlation between CA and standardized K-ABC composite raw score (r = -.16). This apparently reflects a higher likelihood of lower functioning children being chronologically old for their grade placement as a result of the extra year option being used in the school district. The correlation between Gesell DA and K-ABC composite raw score was .59.

Since teacher judgement was used as the criterion of whether children were ready for first grade, it is instructive to know what teachers in the study consider readiness to be. An analysis of the teacher ratings made it possible to compare the relative weighting of developmental maturity and of cognitive skills in teachers' judgements of readiness. The ratings for social maturity and learning habits, which were highly correlated (r = .72), were added together to comprise a "developmental maturity" factor score. Similarly, the highly correlated ratings for language skills and pre-academic skills (r = .86) were summed to comprise a "cognitive skills" factor. On the surface, teachers weighted developmental maturity and cognitive skills equally in judging readiness. For the entire kindergarten group, the overall readiness ratings were similarly correlated with developmental maturity (r = .53) and with cognitive skills (r = .56). Looking at the patterns for individual teachers, however, revealed some variation: four of the teachers weighted developmental maturity more heavily in rating students' readiness, two teachers weighted cognitive skills more heavily, and two teachers gave comparable weight to both factors. The developmental maturity factor and ratings of readiness were highly correlated (ie., correlations above .80) for five of the nine teachers. Thus, we find that both cognitive and non-cognitive aspects of developmental functioning figured significantly in teachers' judgements about readiness.

The developmental maturity and cognitive skills dimensions derived from teacher ratings were used to investigate whether the GSRST "essentially measures the same thing that IQ tests measure" (Shepard & Smith, 1987) or whether it accomplishes the objective of measuring developmental functioning, more broadly defined. Table 2 presents a discriminant analysis comparing validity coefficients for the GSRST and the K-ABC composite. With teacher ratings as the criterion, the K-ABC composite score was more highly correlated with cognitive skills than with developmental maturity, while

the GSRST correlated with the developmental maturity factor only. A similar but less striking pattern is found in comparing the two measures' correlations with parent reports from the Self-Help/Social Questionnaire. The K-ABC correlated more highly with the Communications and Motor Skills domains, while the GSRST was more highly correlated with Daily Living Skills.

Insert Table 2 about here

A concurrent validity analysis addressed the question of whether developmental examiners, based on their administration of the GSRST, arrived at the same conclusions about children's readiness for first grade as kindergarten classroom teachers. Agreement rates were determined by using the school district's cutoff for placement recommendations: a Gesell DA of 5-1/2 or under in the spring of the kindergarten year is interpreted as a recommendation for kindergarten or pre-first grade, rather than first grade. Test outcome and teacher judgement were regarded as a match when a Gesell DA indicating unreadiness for first grade corresponded with the kindergarten teacher's rating of ". . .will definitely not be ready for first grade" or ". . .will probably require special educational services."

In marked contrast to the assertion that "school readiness tests all have error rates in the range of 50%" (Peck, McCaig, & Sapp, 1988), the developmental examiners and classroom teachers in this study agreed at an 81 percent rate (39 of 48 cases) in their readiness recommendations (see Table 3). The hit rate pattern is also striking

---

Insert Table 3 about here

---

in that a markedly high proportion of children were judged unready by the GSRST (58 percent of the sample) and by kindergarten teachers (69 percent of the sample). In fact, so few children in this sample were rated by kindergarten teachers as "definitely ready" for first grade that it was necessary to include the "may not be ready" group (labeled "possibly ready" in Table 3) with the "definitely ready" group in order to comprise a criterion group of reasonable size. Of the nine disagreements between test results and teacher ratings, seven were instances in which the child was judged unready for first grade by the teacher but recommended for first grade on the basis of GSRST results.

Could the high incidence of "unreadiness" be an artifact resulting from the atypical composition of the sample (ie., high in minority group representation and bilinguality)? In actuality, the sample is not particularly low functioning, as indicated by performance on the K-ABC. The sample performed at age expectations on the two K-ABC subtests administered that have a relatively low verbal component (Gestalt Closure and Triangles). In contrast, the sample appeared subnormal on the GSRST (which, it may be noted, has limited verbal content): 42 of the 48 children obtained developmental age scores below their chronological ages, and 32 of 48 had a DA/CA discrepancy of 6 months or greater.

Further analyses were conducted to determine whether the GSRST predicted equally well for different subgroups of children. The sample was divided approximately in half by age, gender, socioeconomic status, language status, and ethnic background. As shown in Table 4, agreement rates between the GSRST and teacher ratings of readiness were not significantly different for subgroups varying by age, SES, and ethnicity. Agreement was significantly higher for females than for males, and significantly lower for bilingual than for primarily English-speaking children.

Insert Table 4 about here

How reliable are the clinically-derived GSRST results? This was addressed by test-retest and inter-rater reliability analyses. To assess test-retest reliability, the GSRST was readministered by a second examiner after an interval of one to five weeks. The mean increase in Gesell DA upon retesting was .9 months. With complete records available for 47 of the children in the sample, there was agreement in 37 of 47 cases, or 79 percent, between the placement recommendations (again, using a DA of 5-1/2 as the cutoff) yielded by the first and second examinations: 26 agreements that a child was unready for first grade and 11 agreements that a child was ready for first grade. The agreement rate was similar when the two examiners were both from the same school district (thus participating in shared training and scoring review experiences) and when the examiners were from different school districts. Test-retest reliability was also computed using correlational analysis which, although failing to address decision-making consistency, controls for the improvement of scores upon retesting. Original and retest scores for 47 children correlated .74.

It was possible to compare this figure of .74 with the reliability coefficient based on quantitative scoring of the GSRST. An item-scoring procedure has been developed for the recently revised Gesell Preschool Test (Walker, 1988), of which the Adaptive Behavior section is essentially the same as the GSRST. Two research assistants with no background in child development were trained to score completed GSRST examiners' records. The test-retest correlation for the GSRST as quantitatively scored by the research assistants was .82. (Note: no classificational agreement rate could be

computed in the absence of any basis for translating the quantitative scores into grade placement recommendations.)

Inter-rater reliability was analyzed using both classificational and correlational measures of agreement. The test records for the second (ie., retest) examinations were independently scored by an experienced trainer of developmental examiners from the Gesell Institute professional staff. Agreement on placement decisions based on GSRST developmental ages obtained by the developmental examiners and by the expert examiner occurred in 43 of 48 cases, with 24 agreements on a recommendation of not ready and 19 agreements on a recommendation of ready. Intraclass correlation coefficients, expressing the strength of association between pairs of score for the same set of test records, were used to directly compare inter-rater reliability for the two scoring systems. The intraclass correlation was .69 for Gesell DA's assigned by developmental examiners, and .92 for the quantitative scores derived by the research assistants.

Discussion and Further Analysis

What are the implications of these findings with respect to appropriate use of the GSRST? The first matter at hand concerns the correspondence of the GSRST with traditional measures of IQ. The correlation of .59 between the Gesell test and K-ABC composite is a close match with the figure of .61 obtained by Kaufman and Kaufman (1972) for concurrent validity of a research version of the Gesell test measured against the Lorge-Thorndike Intelligence Test. This correlation is also of the same magnitude as correlations between the K-ABC and other cognitive tests (Kaufman & Kaufman, 1983). This does not necessarily signify, however, that the GSRST is measuring intelligence. As Wechsler (1974) points out, performance on so-called intelligence tests is dependent upon a number of "nonintellective factors"--drives, attitudes, and

values that manifest in "such traits as persistence, zest, impulse control, and goal awareness" (p. 6). Two measures that correlate in the neighborhood of .6 have approximately one-third total variance in common. It is quite possible, especially for younger children, that most of the common variance between two mental measures can be accounted for by adaptive and developmental factors that are independent of what is narrowly conceived of as "IQ".

Consistent with this interpretation, the discriminant analysis comparing GSRST and K-ABC validity coefficients with teacher ratings (Table 2) suggests that the GSRST is more a measure of developmental maturity than of cognitive ability. Admittedly, the term "developmental maturity" is subject to interpretation. Is developmental maturity as operationally defined in this study (ie., as a composite of teacher ratings of learning habits and social maturity) the equivalent of "adaptive behavior"-- the term used by Gesell (1940) to refer to a child's capacity to apply the developmental skills required by the Gesell examination tasks? Gesell's definition of adaptive behavior, combining instinctive behavior and experience-mediated "intelligent responses," is so inclusive that characterizing test behavior as such would serve little function. As for adaptive behavior as it is commonly defined (ie., in terms of performing daily activities or meeting environmental demands), the correlations between the GSRST and the Self-Help/Social Questionnaire suggest that this may be an aspect of what the test measures, but not a primary one. Given the limited extent to which this study was able to determine what the Gesell test <u>does</u> measure, it remains a moot question as to whether placing children on the basis of school readiness testing is tantamount to ability tracking.

How accurately does the GSRST identify children who, in the judgement of the kindergarten classroom teachers, are unready to progress to the next grade level? The correspondence between readiness recommendations generated by developmental

examiners using the GSRST and by kindergarten teachers observing a child in the classroom over a school year was fairly high, over 80 percent. Both the test and the teachers are evidently attending to a similar constellation of behaviors. The critical question is whether this constellation consists of a broad set of educationally-relevant behaviors, or a specific set of test-related tasks that teachers know to look for as a consequence of their Gesell training.

Another striking finding is the high proportion of children identified as unready by the GSRST. One might suspect that this is a reflection of the sample being atypical. Although clearly not nationally representative given the high proportion of minority and bilingual children, the sample performed within average range on the K-ABC and fully at age expectations on the two K-ABC subtests assessing nonverbal cognitive tasks (the content most directly comparable to the GSRST). It should be also noted that over 40 percent of the children in the higher SES and the white subgroups were identified as unready (see Table 4). The finding that a disproportionate number of children receive developmental age scores below their chronological age is highly consistent with recent research (Walker, 1988) on the normative sample of the Gesell Preschool Test (Walker, 1988).

What is particularly noteworthy in this study is that teacher judgement was even more skewed toward identifying children as unready for the next grade than GSRST results. Is this a natural tendency for kindergarten teachers, or might it be fostered by exposure to Gesell assessment practices and philosophy? To investigate this, each of the kindergarten teachers was assigned to one of three distinguishable levels according to degree of Gesell exposure, ranging from attending a single Gesell Institute conference to becoming trained as a qualified developmental examiner. Exposure level was then correlated with teachers' mean readiness rating for their class. The correlation was highly negative ($r = -.74$, $p < .05$), indicating that the

tendency to perceive children as unready is directly proportional to the extent of Gesell Institute training received. Might this statistical association reflect classroom differences in skills or maturity, in that developmentally young children might be systematically assigned to teachers who are most trained to provide for them? A partial correlation controlling for classroom mean differences in teacher ratings of school performance was computed, revealing an even stronger relationship between Gesell training and mean ratings of school readiness ($r = -.91$, $p < .01$). Although a causal relationship is suggested, the possibility of reciprocal effects should be considered, ie., those teachers who tend to perceive children as problematically lacking in school readiness may be the most inclined to pursue professional training in this area.

Of the demographic variables analyzed, agreement rates between the GSRST and teacher ratings of readiness were significantly different for subgroups defined by gender and by language status, but not by age, SES, or ethnicity. The gender effect may be a sampling artifact, as suggested by substantially greater variances for males than for females on all GSRST and K-ABC performance variables. The significantly higher disagreement rate for bilingual children than for primarily English-speaking children despite comparable GSRST performance for the two subgroups is of particular interest. It is possible that some disagreements resulted from teachers weighing educational preparedness, which would expectedly be lower for bilingual children, into their judgements about readiness. They would thus be inclined to rate readiness in a manner inconsistent with the GSRST, which minimizes the influence of language. Although the developmental focus of the GSRST may be advantageous in pursuing the goal of "culture fair" measurement, another perspective is that deemphasizing the language functioning of bilingual children fails to encompass a primary determinant of their educational performance. The finding of greater disagreement between

alternate assessments of readiness for bilingual children simply adds psychometric support to the observation that the school readiness model has serious limitations for diagnosing and addressing the educational needs of bilingual children, and would best be integrated into a system offering specialized (ie., language-focused) services and assessment techniques.

A higher rate of unready recommendations from GSRST testing was found for lower-middle SES children (67%) than for middle/upper-middle SES children (50%), but not to a statistically significant degree. As an illustration of the limited impact of SES upon GSRST performance, it is instructive to compare the GSRST with the K-ABC, which is promoted as being less culturally biased than traditional measures of intelligence. The relative influence of SES upon the two tests was gauged by correlating each with the SES variable used to split the sample into the two SES subgroups, a weighted composite of parent education and income data. The correlation of the K-ABC composite with SES (r = .42) was higher than the correlation of the GSRST with SES (r = .28), although the difference between these correlations was not statistically significant. A similar pattern was found when bilingual children were omitted from the analysis, with resulting correlations of .27 and -.02, respectively, with SES (n = 26, difference not significant at $p < .05$, two-tailed).

How reliable is the GSRST? Test-retest agreement on recommendation decisions in 79 percent of cases and a reliability coefficient of .74 clearly contraindicate using the test as the sole, or even primary, basis for grade placement decisions. More appropriately, the GSRST might be used as one of several sources of information in making developmental grade placement decisions.

Inter-rater reliability, with agreement in 89 percent of cases and a reliability coefficient of .77, also fell below acceptable range for making placement decisions. The figures for inter-rater reliability should be regarded with some caution, as they

are based on a second rater's review of original test records and products rather than actual observation of the child's performance. This finding is more appropriately interpreted as showing that substantial disagreement may be expected between fully qualified examiners independently reviewing the same test records. In fact, Walker (1988) obtained a nearly identical figure ($r = .75$) for the inter-rater agreement between expert examiners independently reviewing records.

As might have been expected, the clinically-based method of test interpretation was less reliable than the quantitative scoring method with regard to both test-retest and inter-rater reliability. The superior reliability of the quantitative scoring method raises the question of whether this approach is to be preferred over the clinically-based scoring method in current use. There are two significant points to consider. First, what works best? Higher reliability is to be expected from an objective quantitative scoring method; there may or may not be a corresponding increment in validity. If the two methods are comparable with respect to validity, however, there is a distinct economic advantage to using the quantitative scoring. This brings up the second point: the time and expense of training school personnel to function as developmental examiners may be beneficial in terms of enhancing educators' understanding of and respect for developmentally appropriate behavior. Thus, a school system placing higher priority on staff training than on cost-effective decision-making might opt to remain with the clinically-based scoring method.

As to which scoring method works best, the absence of follow-up criterion data (ie., for predictive validity analyses) limits this study's contribution to the matter. Still, it is instructive to make what validity comparisons are available. As shown in Table 5, the two scoring methods produced generally comparable validity coefficients, with none of the differences between corresponding correlations reaching statistical significance at $p < .05$.

Insert Table 5 about here

Rather than looking for evidence to favor one approach or another, the optimal solution may involve concurrent use of quantitative and clinical approaches to interpretation as is widely practiced by clinical, school, and neuro-psychologists. While clinical interpretation of process variables (incidental behavior, strategies, attentional and motivational factors, etc.) by a skilled and experienced clinician significantly enhances the value of an individually-administered test, an objectified quantitative scoring system enables a test to be used responsibly by a trained examiner with limited experience in clinical interpretation and provides a point of departure from which the experienced examiner can apply clinical interpretations.

Summary

The construct of school readiness as measured by the GSRST remains in need of clearer definition. Although the GSRST has a moderately high correlation with the K-ABC, a measure of general intellectual functioning, the possibility that non-cognitive factors play a significant part in this relationship cannot be dismissed. Unlike the K-ABC, performance on the GSRST was more highly associated with teacher ratings of social behavior and learning habits than of cognitive skills.

The GSRST clearly validated teacher judgements of unreadiness, as is commonly claimed. But given the excessive rates at which both teacher judgement and test results identified children as unready, the more apt conclusion may be that both mechanisms are in need of "recalibration." The tendency for teachers to perceive children as unready in this study may be a consequence of the Gesell training of

teachers in this school system, as suggested by the strong association demonstrated between these two factors.

Test-retest reliability and inter-rater reliability for the clinically-based scoring method currently is use were well below levels typically regarded as acceptable for making placement decisions. One might conclude that, although the GSRST is not sufficiently reliable and valid to justify using it as the primary determinant of grade placement, it is a valuable source of information about school readiness--whatever that may be.

References

Bear, G.G., & Modlin, P.D. (1987). Gesell's developmental testing: What purpose does it serve? Psychology in the Schools, 24, 40-44.

Bradley, R.H. (1985). Review of The Gesell School Readiness Test. In J. Mitchell, Jr. (Ed.), The ninth mental measurements yearbook. Lincoln, NE: University of Nebraska Press.

California State Department of Education (1988). Here they come: Ready or not! Report of the school readiness task force. Sacramento: Author.

Educational Research Service (1986). Kindergarten programs and practices in public schools. Arlington, VA.: Author.

Gesell, A. (1940). The first five years of life: A guide to the study of the preschool child. New York: Harper & Row.

Gesell Institute (1985). Scoring notes: The developmental examination (rev. ed.). New Haven: Gesell Institute of Human Development.

Gredler, G. (1984). Transition classes: A viable alternative for the at-risk child? Psychology in the Schools, 21, 463-470.

Ilg, F.L., & Ames, L.B. (1964). School readiness: Behavior tests used at the Gesell Institute. New York: Harper & Row.

Kaufman, A.S. (1971). Piaget and Gesell: A psychometric analysis of tests built from their tasks. Child Development, 42, 1341-1360.

Kaufman, N.L. (1985). Review of Gesell Preschool Test. In J. Mitchell, Jr. (Ed.), The ninth mental measurements yearbook. Lincoln, NE: University of Nebraska Press.

Kaufman, A.S., & Kaufman, N.L. (1983). K-ABC: Kaufman Assessment Battery for Children. Circle Pines, MN: American Guidance Service.

Lichtenstein, R., & Ireton, H. (1984). Preschool screening: Identifying young children with developmental and educational problems. Orlando: Grune & Stratton.

Meisels, S. J. (1987). Uses and abuses of developmental screening and school readiness tests. Young Children, 42, 4-6, 68-73.

Meisels, S.J. (in press). High-stakes testing in kindergarten. Educational Leadership

National Association for the Education of Young Children (1988). NAEYC position statement on standardized testing of young children 3 through 8 years of age. Young Children, 43, 42-47.

National Association of Early Childhood Specialists in State Departments of Education (1987). Unacceptable trends in kindergarten entry and placement: A position statement. Lincoln, NE: author.

Peck, J.T., McCaig, G. & Sapp, M.E. (1988). Kindergarten policies: What is best for children? Research Monographs of the National Association for the Education of Young Children, 2

Shepard, L.A., & Smith, M.L. (1985) Boulder Valley kindergarten study: Retention practices and retention effects. Boulder, CO: Boulder Valley Public Schools.

Shepard, L.A., & Smith, M.L. (1986). Synthesis of research on school readiness and kindergarten retention. Educational Leadership, 44, 78-86.

Sparrow, S.S., & Cicchetti, D.V. (in press). AGS Early Screening Profiles: Self-Help/Social Questionnaire. Circle Pines, MN: American Guidance Service.

Walker, R.N. (1988). The Gesell screening examinations: Psychometric properties. Manuscript submitted for publication.

Wechsler, D. (1974). Manual for the Wechsler Intelligence Scale for Children - Revised. San Antonio: Psychological Corporation.

Table 1

Descriptive statistics for the Gesell School Readiness Screening Test and Kaufman Assessment Battery for Children

		Gesell DA/CA discrepancy (mos.)			K-ABC composite scaled score		
Sample	n	Mean	S.D.	F	Mean	S.D.	F
Total group	48	-7.6	6.2		9.9	2.4	
By age							
Five year olds	24	-4.8	6.0	12.78**	10.3	2.6	1.08
Six and over	24	-10.5	5.1		9.5	2.2	
By gender							
Males	24	-8.3	7.2	0.62	10.5	2.7	3.49
Females	24	-6.9	5.1		9.3	2.2	
By SES level							
Middle/upper-middle	24	-6.4	5.8	1.97	10.6	2.2	4.72*
Lower-middle	24	-8.9	5.7		9.2	2.4	
By language status							
English-speaking	26	-7.5	6.6	0.04	10.9	2.4	12.00**
Bilingual	22	-7.8	5.9		8.7	1.8	
By ethnic background							
White	26	-6.3	6.6		10.7	2.4	
Hispanic	16	-10.0	5.9	1.81	8.5	1.9	5.17**
Black	5	-7.4	4.6		10.6	2.4	

* $p < .05$ ** $p < .01$

Table 2

Correlations Comparing the Gesell School Readiness Screening Test and Kaufman Assessment Battery for Children

	Gesell DA	K-ABC composite raw score
Composite teacher ratings[a]		
Developmental maturity	.40**	.21
Cognitive skills	.01	.36*
Self-Help/Social Questionnaire[b]		
Communication	.31*	.44**
Daily Living Skills	.14	.00
Socialization	.22	.28
Motor Skills	.22	.30*
Total score	.30*	.34*
Chronological age[b]	.32*	.19

[a] N = 46. [b] N = 48.

* p < .05. ** p < .01.

Table 3

Agreement Between Placement Recommendations Based on Gesell Testing and Teacher Judgement

	Teacher ratings of first grade readiness			
Recommendation from Gesell testing	Special needs	Unready	Possibly ready	Ready
Kindergarten or pre-first grade	7	19	0	2
First grade	2	5	6	7

Table 4

Comparisons of Agreement Rates Between Gesell Testing and Teacher Judgement

	First grade recommendations					
	Gesell-no Tchr.-no	Gesell-no Tchr.-yes	Gesell-yes Tchr.-no	Gesell-yes Tchr.-yes	Agreements/ Disagreements	χ^2 [a]
By age						
Five year olds	16	0	4	4	20/4	0.30
Six and older	10	2	3	9	19/5	
By sex						
Male	11	1	6	6	17/7	4.08*
Female	15	1	1	7	22/2	
By SES						
Middle/upper-mid	12	0	4	8	20/4	0.30
Lower-middle	14	2	3	5	19/5	
By language status						
English-speaking	12	1	2	11	23/3	3.90*
Bilingual	14	1	5	2	16/6	
By ethnic background [b]						
White	10	1	5	10	20/6	1.47
Non-white	16	1	1	3	19/2	

[a] Using Yates' correction for continuity. df = 1.

[b] The one Oriental student was omitted from this analysis.

* $p < .05$.

Table 5

Concurrent validity of clinically-based and quantitative scoring methods for the Gesell School Readiness Screening Test

Criterion measures	Clinically-based (conventional)	Quantitative
Teacher rating factors[a]		
Developmental maturity	.40**	.46**
Cognitive skills	.01	.11
Self-Help/Social Questionnaire[b]		
Communications	.31*	.33*
Daily Living Skills	.14	.10
Socialization	.22	.17
Motor Skills	.22	.27
ABP total score	.30*	.29*
K-ABC scaled scores[b]		
Composite	.47**	.52**
Simultaneous processing (non-verbal subtests)	.46**	.55**

[a] N = 46 [b] N = 48

* $p < .05$ ** $p < .01$